The GreenCut2 Resource: A phylogenomically-derived inventory of proteins specific to the plant lineage

Steven J. Karpowicz, Simon E. Prochnik, Arthur R. Grossman, and Sabeeha S. Merchant

Table S1. Cyanobacterial genomes searched for homologs of GreenCut2 proteins. Arabidopsis GreenCut2 orthologs were used in BLASTP searches to identify homologs encoded by 37 cyanobacterial genomes. CyOG denotes a genome used in a previous analysis of conserved cyanobacterial proteins (2). Proteins denotes the number of predicted proteins in an organism's NCBI RefSeq dataset. Habitat: T=Terrestrial, M=Marine, FW=Fresh Water, HS=Hot Spring (Thermophilic). Phenotype: NF=Nitrogen Fixing, Fil=Filamentous, HL=High light adapted, LL=Low light adapted, PCB=Prochlorococcus chlorophyll binding proteins, PBS=Phycobilisomes.

Table S2. Proteins of the GreenCut2 and protein functional features. The JGI protein identifier numbers of the Chlamydomonas orthologs are shown in column A. The Chlamydomonas gene names and gene definitions are given in columns B and C, followed by a suggested protein function for unknown proteins in column D. The Arabidopsis ortholog is given in column E as a hyperlink to the relevant TAIR web page. If more than one Arabidopsis ortholog was present for a Chlamydomonas protein, then the additional co-ortholog(s) are listed immediately below the first co-ortholog. Column F lists whether the protein is known, known with inferred function, unknown, or unknown with predicted function. Column G lists if experimental data or only software prediction supported localization for a protein. Column H lists the location of the protein in the cell. A domain as predicted by Pfam is shown in column I. The classification of the Arabidopsis ortholog by MapMan is given in column J. Column K presents the Arabidopsis organ in which the encoding transcript was dominantly expressed. The number of cyanobacterial genomes encoding a homolog to the Arabidopsis protein is shown in column L. Column M lists the functional category assigned to the GreenCut2 orthologs as determined by this study. Column N indicates the insertion site of a T-DNA as indicated on the TAIR website. Columns O and P list, respectively, whether a knockout strain of the Arabidopsis ortholog is present in the Chloroplast 2010 project collection, and what is the measured phenotype of the mutant (1). Column Q indicates whether the GreenCut2 ortholog may be a false positive in this protein inventory. Column R contains Pubmed IDs for relevant publications for those proteins that were classified as unknown function in the version 1 GreenCut2 but which are now classified as known function. A description of placeholder names for genes of unknown function (CPLD, CGLD, CGL, CPL) is given at the bottom of the table. These gene names are not intended to supersede any gene names given in the literature.

Table S3. GreenCut2 proteins not encoded on additional genomes. Results from the genomes of plants, algae, or a diatom are given in separate sheets. Columns A-F are the same as in **Table S2**. An 'X' in column G/H/I indicates that the protein was missing from the organism listed on line two of the respective column.

Table S4. Potential false positives in the GreenCut2. Columns A-O in the 'false positives' sheet are the same as in **Table S2**. Column P indicates the reason why a GreenCut2 protein was identified as a potential false positive. BLASTP searches against the NCBI nr database using both the Chlamydomonas and Arabidopsis GreenCut2 orthologs were conducted to observe if non-plant orthologs were found.

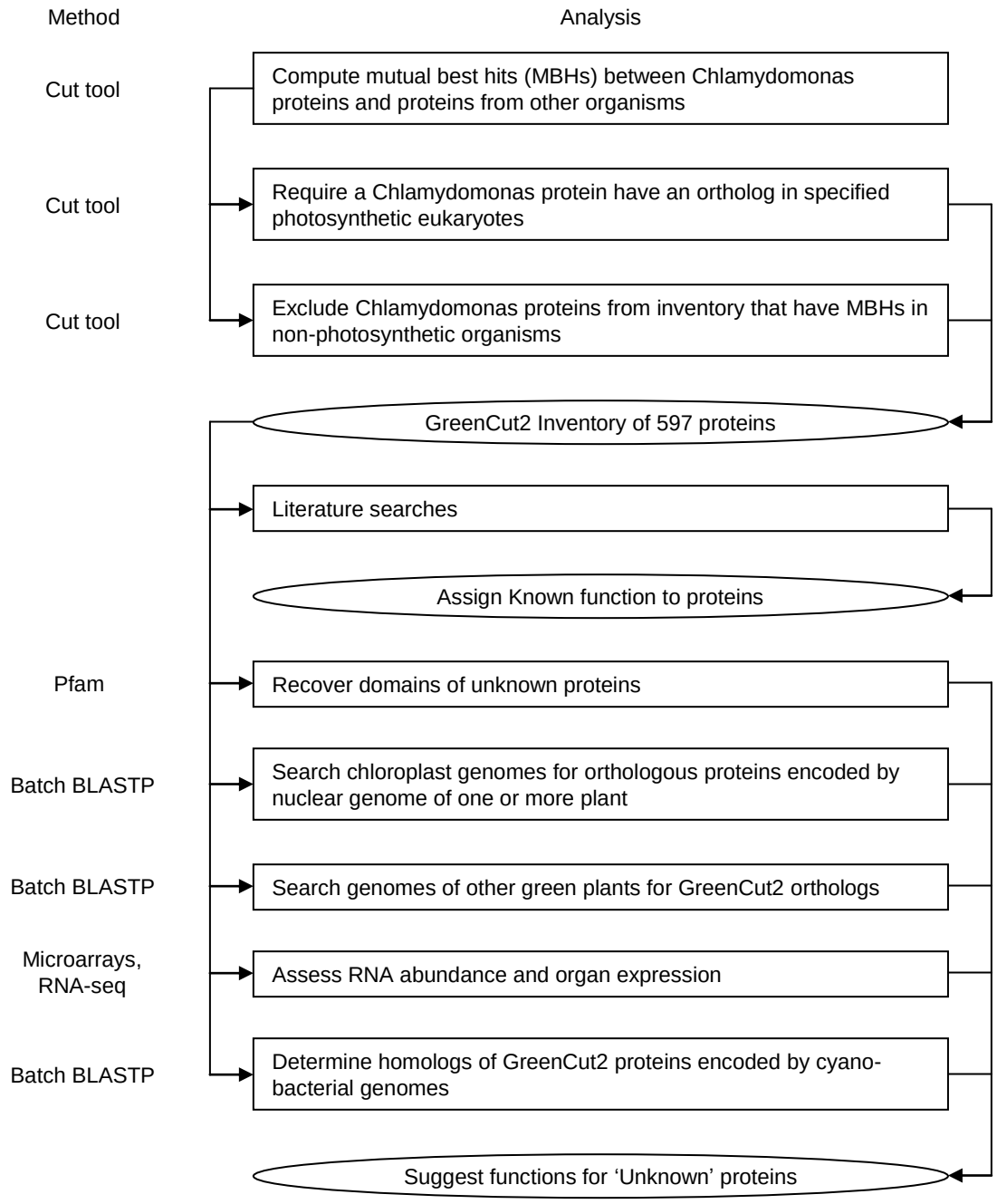

Figure S1. Workflow for generation of the GreenCut2 and analysis of GreenCut2 proteins.

For details of each step, see METHODS.

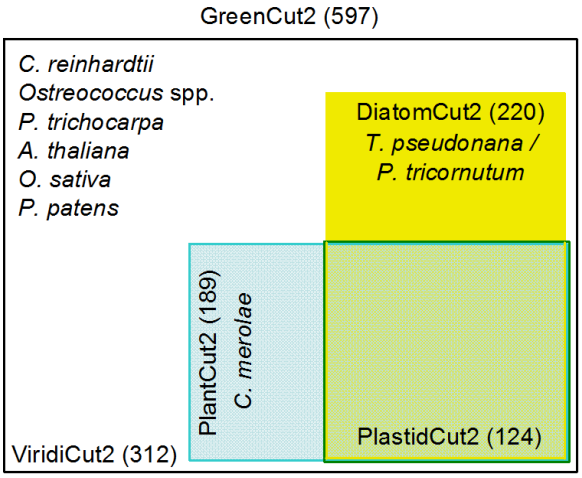

Figure S2. Diagram of the sub-groups of the GreenCut2. The name of the sub-group and organisms with the given number of orthologs in the sub-group are shown within each box. The GreenCut2 (black square) includes all proteins with orthologs in the organisms listed in the upper left corner. The PlantCut2 (blue square and rightward lines) and DiatomCut2 (yellow square and leftward lines) are sub-groups of the GreenCut2. The proteins shared by green plants, a red alga, and diatoms is defined as the PlastidCut2, which is marked by a green square. All proteins of the GreenCut2 that are not conserved in a red alga or diatoms belong the ViridiCut2 sub-group.

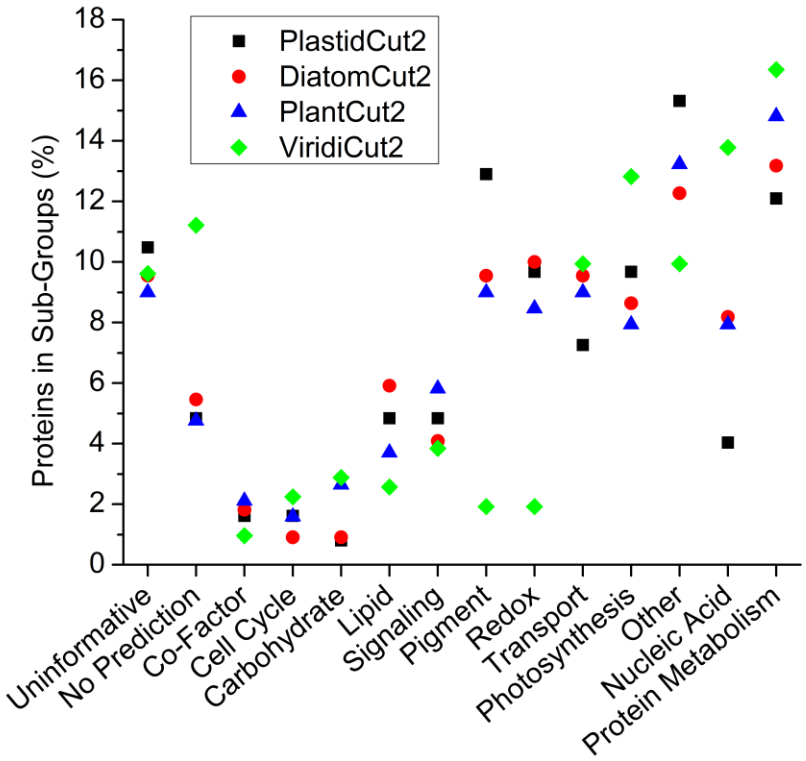

Figure S3. Percentage of proteins of each functional group within each sub-group of the GreenCut2. The number of functional category proteins within each sub-group was scaled by the total number of proteins within a sub-group. Functional categories are as described in Figure 2.

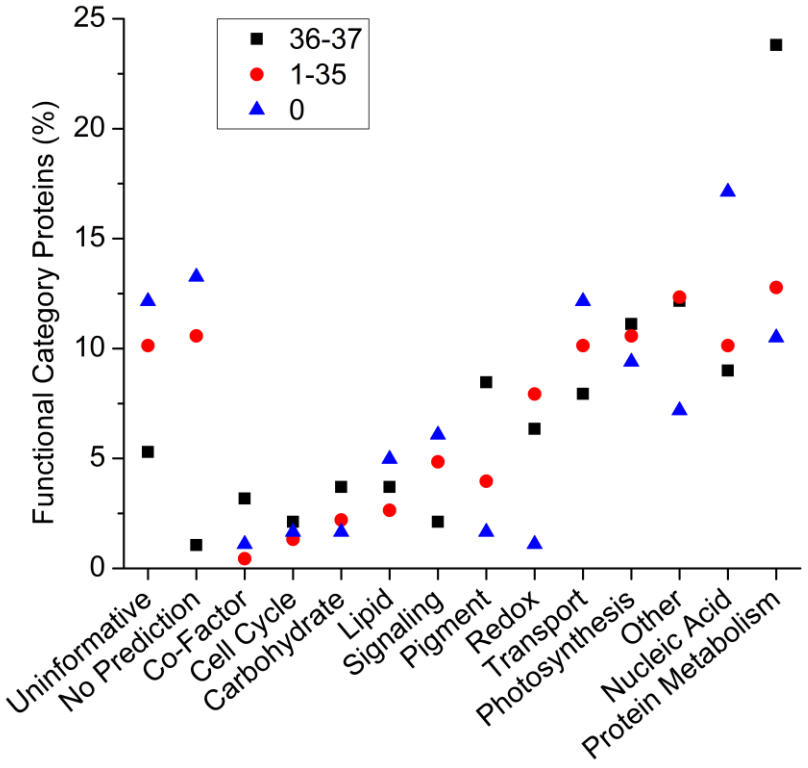

Figure S4. Percentage of proteins of each functional group that have cyanobacterial homologs. The number of functional category proteins within each sub-group was scaled by the total number of proteins within a sub-group. Functional categories are as described in Figure 2. Each data point represents the number of GreenCut2 proteins that had a homolog in the given number of cyanobacterial genomes.

Table S1. Cyanobacterial genomes searched for homologs of GreenCut2 proteins

Organism	CyOG	Proteins	Habitat				Phenotype					
			T	M	FW	HS	NF	Fil	HL	LL	PCB	PBS
Acaryochloris marina MBIC11017		8409		X								X
Anabaena variabilis ATCC 29413	X	5661	X		X		X	X				X
Cyanothece sp. ATCC 51142		5304		X			X					X
Cyanothece sp. PCC 7424		5880			X		X					X
Cyanothece sp. PCC 7425		5428			X		X					X
Cyanothece sp. PCC 8801		4566			X		X					X
Cyanothece sp. PCC 8802		4648			X		X					X
Gloeobacter violaceus PCC 7421	X	4430	X		X							X
Microcystis aeruginosa NIES-843		6312			X							X
Nostoc sp. PCC 7120	X	6310	X		X		X	X				X
Nostoc punctiforme PCC 73102		6690	X		X		X	X				X
Prochlorococcus marinus str. AS9601		1921		X					X		X	
Prochlorococcus marinus str. MIT 9211		1855		X						X	X	
Prochlorococcus marinus str. MIT 9215		1983		X					X		X	
Prochlorococcus marinus str. MIT 9301		1907		X					X		X	
Prochlorococcus marinus str. MIT 9303		2997		X						X	X	
Prochlorococcus marinus str. MIT 9312	X	1810		X					X		X	
Prochlorococcus marinus str. MIT 9313	X	2269		X						X	X	
Prochlorococcus marinus str. MIT 9515		1906		X					X		X	
Prochlorococcus marinus str. NATL1A		2193		X						X	X	
Prochlorococcus marinus str. NATL2A	X	2163		X						X	X	
Prochlorococcus marinus subsp. *marinus* str. CCMP1375 (SS120)	X	1883		X						X	X	
Prochlorococcus marinus subsp. *pastoris* str. CCMP1986 (MED4)	X	1717		X					X		X	
Synechococcus elongatus PCC 6301	X	2527			X							X
Synechococcus elongatus PCC 7942	X	2662			X							X
Synechococcus sp. CC9311		2892		X						X		X
Synechococcus sp. CC9605	X	2645		X						X		X
Synechococcus sp. CC9902	X	2307		X								X
Synechococcus sp. JA-2-3B'a(2-13)		2862			X	X	X		X			X
Synechococcus sp. JA-3-3Ab		2760			X	X	X		X			X
Synechococcus sp. PCC 7002		3186		X				X	X			X
Synechococcus sp. RCC307		2535		X					X			X
Synechococcus sp. WH 7803		2533		X								X
Synechococcus sp. WH 8102	X	2519		X								X
Synechocystis sp. PCC 6803	X	3569			X							X
Thermosynechococcus elongatus BP-1	X	2476			X	X						X
Trichodesmium erythraeum IMS101		4451		X			X	X				X

Literature Cited

1. Ajjawi, I., Lu, Y., Savage, L. J., Bell, S. M., and Last, R. L. (2010) *Plant Physiol.* **152**, 529-540
2. Mulkidjanian, A. Y., Koonin, E. V., Makarova, K. S., Mekhedov, S. L., Sorokin, A., Wolf, Y. I., Dufresne, A., Partensky, F., Burd, H., Kaznadzey, D., Haselkorn, R., and Galperin, M. Y. (2006) *Proc. Natl. Acad. Sci. USA* **103**, 13126-13131